# Text-Image Separation in Document Images using Boundary/Perimeter Detection

Priti P. Rege[1], Chanchal A. Chandrakar[2]
College of Engineering,, Pune, India.
Email: [1] ppr.extc@coep.ac.in
[2] chanch.chandra@gmail.com

***Abstract*—Document analysis plays an important role in office automation, especially in intelligent signal processing. The proposed system consists of two modules: block segmentation and block identification. In this approach, first a document is segmented into several non-overlapping blocks by utilizing a novel recursive segmentation technique, and then extracts the features embedded in each segmented block are extracted. Two kinds of features, connected components and image boundary/perimeter features are extracted. Document with text inside image pose limitations in earlier reported literature. This is taken care of by applying additional pass of the Run Length Smearing on the extracted image that contains text. Proposed scheme is independent of type and language of the document.**

***Index Terms*— OCR, Segmantation, Connected components.**

## I. Introduction

Work in the field of document image processing covers many different areas including preprocessing, layout analysis, optical character recognition, graphics analysis, form processing, and writer identification. Text parts are the main information carriers in most applications. For that purpose, it is necessary to locate text objects within the image, recognize them, and extract the hidden information. The documents may contain, besides text, graphics and images that overlap. Since text lines are not always horizontally aligned, finding text parts and locating characters, words, and lines are not trivial tasks. Due to the tremendous reduction in the storage space of the processed results, it is advantageous to reproduce, transmit, and store the document in the processed form. The extracted regions can then be processed by a subsequent step according to their types, e.g. OCR for text regions and compression for graphics and halftone images.

Several methods have been proposed to separate text and image or text and graphics from the background. Two broad categories of available methods are Connected Component based (CC) and texture based algorithms. The first category segments the image into a set of CCs and then classifies each CC as either text or non text. CC based algorithms are relatively simple The underlying assumption is that texts in document images can be seen as sets of separating connected components each of which has distinct intensity or color distribution and linked edge contours. Up to now several strategies have been tried to solve the problem of segmentation. Techniques for page segmentation and layout analysis are broadly divided in to three main categories: top-down, bottom-up and hybrid techniques [1]. Many bottom-up Approaches are used for page segmentation and block identification [5], [7]. Yuan, Tan [2] designed method that makes use of edge information to extract textual blocks from gray scale document images. It aims at detecting only textual regions on heavy noise infected newspaper images and separate them from graphical regions. The White Tiles Approach [3] described new approaches to page segmentation and classification. In this method, once the white tiles of each region have been gathered together and their total area is estimated, and regions are classified as text or images. George Nagy, Mukkai Krishnamoorthy [4] have proposed two complementary methods for characterizing the spatial structure of digitized technical documents and labeling various logical components without using optical character recognition. Projection profile method [6], [8] is used for separating the text and images, which is only suitable for Devanagari Documents (Hindi document). The main disadvantage of this method is that the irregular shaped images with non-rectangular shaped text blocks may result in loss of some text. They can be dealt with by adapting algorithms available for Roman script. Kuo-Chin Fan, Chi-Hwa Liu, Yuan-Kai Wang [9] have implemented a feature-based document analysis system which utilizes domain knowledge to segment and classify mixed text/graphics/image documents. This method is only suitable for pure text or image document, i.e. a document which has only text region or image region. This method is good for text-image identification not for extraction. The Constrained Run-Length Algorithm (CRLA) [10] is a well-known technique for page segmentation. The algorithm is very efficient for partitioning documents with Manhattan layouts but not suited to deal with complex layout pages, e.g. irregular graphics embedded in a text paragraph. Its main drawback is the use of only local information during the smearing stage, which may lead to erroneous linkage of text and graphics. Kuo-Chin Fan, Liang-Shen Wang, Yuan-Kai Wang [11] proposed an intelligent document analysis system to achieve the document segmentation and identification goal. The proposed system consists of two modules: block segmentation and block identification. Two kinds of features, connectivity histogram and multi resolution features are extracted.

## II. Methodology

### *A. Document Segmentation*

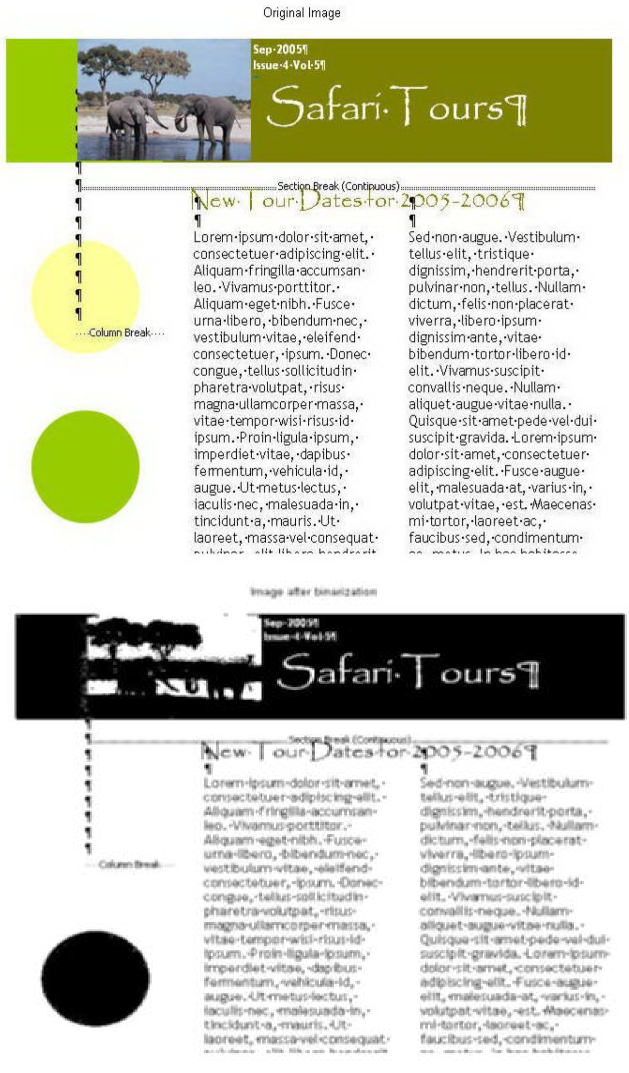
The main aim of document segmentation is to segment a document into several separate blocks with each block

DOI: 01.IJSIP.4.1.1044

representing one type of medium. The method we adopted in segmenting documents is a combination of run-length smearing algorithm and boundary perimeter detection procedure. It first performs the smearing operation in both the horizontal and vertical directions to generate blocks. Due to the selection of an improper threshold in the run-length smearing process, an intact paragraph might be segmented into several consecutive horizontal stripes with each stripe representing one horizontal text line. The stripe merging procedure is thus devised to merge those text stripes that belong to the same paragraph. The merit is that it generates an intact text block instead of several smaller text stripes. By this way, it not only saves storage space but also reduces the burden in performing optical character recognition.

### *B. Pre-processing*

Preprocessing of document images is the way of using mature image processing techniques to improve the quality of images. Its purpose is to enhance and extract useful information of images for later processing purposes. Two preprocessing tasks, thresholding and noise removal, are performed here.



Lorem·ipsum·dolor·sit·amet,· consectetuer·adipiscing·elit.· Aliquam·fringilla·accumsan· leo.·Vivamus·porttitor.· Aliquam·eget·nibh.·Fusce· urna·libero,·bibendum·nec,· vestibulum·vitae,·eleifend· consectetuer,·ipsum.·Donec· congue,·tellus·sollicitudin· pharetra·volutpat,·risus· magna·ullamcorper·massa,· vitae·tempor·wisi·risus·id· ipsum.·Proin·ligula·ipsum,· imperdiet·vitae,·dapibus· fermentum,·vehicula·id,· augue.·Ut·metus·lectus,· iaculis·nec,·malesuada·in,· tincidunt·a,·mauris.·Ut· laoreet,·massa·vel·consequat·

Sed·non·augue.·Vestibulum· tellus·elit,·tristique· dignissim,·hendrerit·porta,· pulvinar·non,·tellus.·Nullam· dictum,·felis·non·placerat· viverra,·libero·ipsum· dignissim·ante,·vitae· bibendum·tortor·libero·id· elit.·Vivamus·suscipit· convallis·neque.·Nullam· aliquet·augue·vitae·nulla.· Quisque·sit·amet·pede·vel·dui· suscipit·gravida.·Lorem·ipsum· dolor·sit·amet,·consectetuer· adipiscing·elit.·Fusce·augue· elit,·malesuada·at,·varius·in,· volutpat·vitae,·est.·Maecenas· mi·tortor,·laoreet·ac,· faucibus·sed,·condimentum·

Figure 1. Original color image and Binarized image

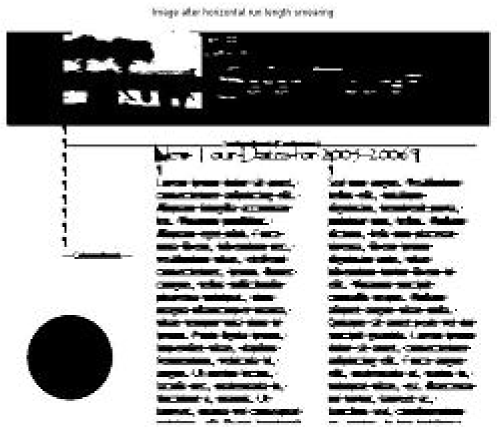


Image after horizontal run length smearing
New·Tour·Dates·for·2005-2006¶

Figure 2. Image after horizontal run length smearing

Document image is usually captured as colored image (RGB) and we convert it into Gray level image. image. Hence, a binarization procedure converting a gray-level image into a binary image is necessary. OTSU Document image is usually captured as colored image (RGB) and we convert it into Gray level image. Thresholding is utilized to accomplish the conversion task. Figure 1, shows the original and binarized image.

### *C. Block Extraction*

With the pre-processing being done, a binarized image is then obtained. In this binarized image, each meaningful block which can be easily recognized by human beings is composed of pixels. RLSA is an operation to connect two nonadjacent runs into one merged run if the distance between these two runs is smaller than a threshold.

It is assumed that white pixels are represented by 0 and black pixels by 1. If the number of 0 between l's is less than or equal to a constant C, then 0 is replaced by 1. In other words, two runs having distance smaller than the threshold C will be merged into one run. Example

Before smearing:

$1111111000001111111\underline{000}11$

After smearing:

$1111111000001111111\underline{111}11$

The result generated by the smearing operation is several l-runs in each horizontal row.

Two constants CHT (Constant Horizontal threshold value) and CVT (Constant Vertical threshold value) are needed due to the processing of Run Length Smearing algorithm (RLSA) in two directions. CHT selected as half of CVT. Figure 2 and 3 illustrate the operation of RLSA in two directions i.e. horizontal and vertical directions. The results generated by horizontal and vertical smearing operations are then combined by logical AND operation to produce the meaningful blocks. In other words, a pixel in the resulting image is black if the pixels at the same position of both vertical and horizontal Run-length smearing images are black. Though run length smearing can combine related pixels into meaningful blocks, there are still some small blocks which should be further combined. We repeatedly apply run length smearing with increased runlengths to the output of AND operation to merge these blochs. An image illustrating the result generated by

DOI: 01.IJSIP.4.1.1044

RLSA is shown in Figure 4. Result of next step, to detect boundary after reapplying horizontal run length smearing, is shown in Figure 5.

### *D. Boundary/Perimeter Detection*

Since we have obtained meaningful blocks, the next step is to find the outside boundary of smearing blocks. Outside boundary is defined as the white pixels surrounding the extracted boundary of a block. The traditional edge detection methods are not adequate for detecting this kind of boundary. For example, if traditional edge detection methods are operated on a one-pixel-width line block open contour will be generated.

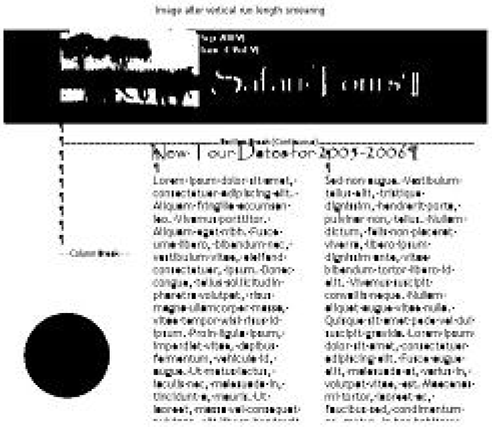


Image after vertical run length smearing
Safari Tours
New Tour Dates for 2005-2006

Figure 3. Image after vertical run length smearing

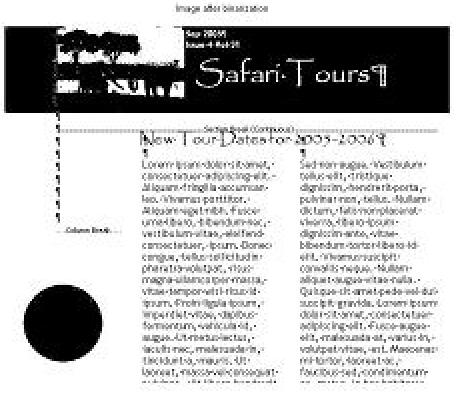


Figure 4. Final Result of Run length smearing

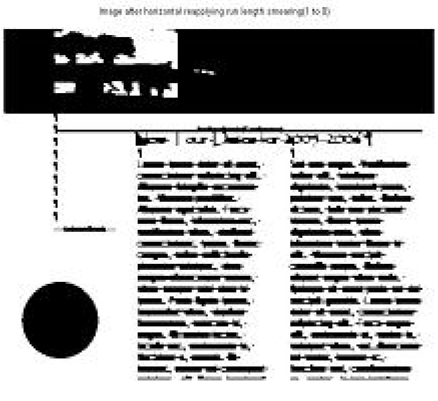


Image after horizontal reapplying run length smearing(1 to 8)
New Tour Dates for 2005-2006

Figure 5. Image after reapplying horizontal run length smearing

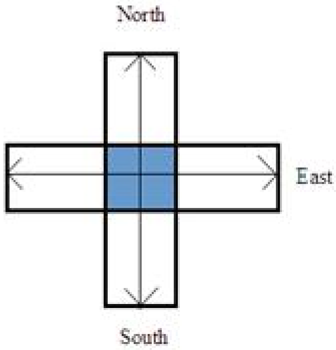


Figure 6. Four possible search Direction

We have used 4-connectivity as well as 8-connectivity for boundary extraction. It is observed that 4-connectivity We have used 4-connectivity as well as 8-connectivity for boundary extraction. It is observed that 4-connectivity gives batter results. Figure 6 shows the four possible search directions. A pixel is considered to be a part of the perimeter, if it is nonzero and it is connected to at least one nonzero value.

Connectivity can also be defined in a more general way for any dimension by using connectivity for a 3-by-3 matrix of 0's and 1's. The 1-valued elements define neighborhood locations relative to the center element of connectivity. Note that connectivity must be symmetric about its center element. Figure 7 shows the image after boundary detection.

### *E. Connected Component and Area Computation*

Connected components are rectangular boxes bounding together regions of connected black pixels. The objective of CC stage is to form rectangles around distinct components in the page. The algorithm used is a simple iterative procedure with successive scan lines of an image to determine whether black pixels in any pair of scan lines are connected together. Bounding rectangles are extended to enclose any groupings of connected black pixels between successive scan lines.

A bounding box stops growing in size only when there are no more black pixels on the current scan line join black pixels from previous line. These bounding boxes then form the skeleton for further analysis. After detection of boundary from the image, we find the connected component. Then, with the help of connected component we compute the area of

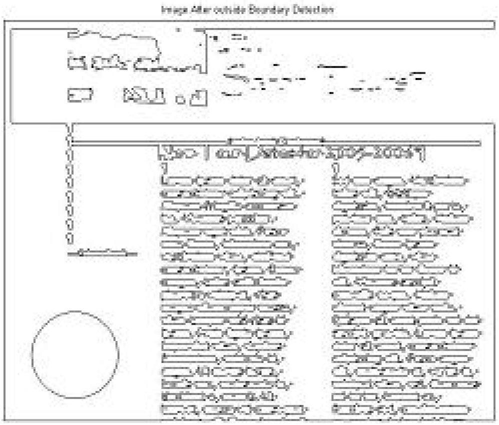


Image After outside Boundary Detection
New Tour Dates for 2005-2006

Figure 7. Image after Boundary Detection

© 2013 ACEEE
DOI: 01.IJSIP.4.1.1044

closed loop. From Figure 6 it is clear that the text which is converted into uniform stripes has uniform area. When we compare the area of the closed loop (stripes which form by run length smearing) with open loop area then, it fills Zero in smaller area. Figure 7 shows the smaller area which is filled with zero or black pixel.

To find the connected component the algorithm uses the following general procedure:

1. Scan all image pixels, assigning preliminary labels to nonzero pixels and recording label equivalences in a union-find table.
2. Resolve the equivalence classes using the union-find algorithm.
3. Relabel the pixels based on the resolved equivalence classes.

### *F. Image-Text Separation*

The text areas are thus filled with zero or black pixel, which is shown in Figure 8. By ORing the Original binary image with image so obtained entire text is separated from the original images. This is shown in Figure 9. After extracting text, we XOR this image with the image in Figure 3 to separated image/graphics in the original binary document. The separated image is shown in Figure 10. Results for two more samples are given in Figure 11 and Figure 12. Text and image parts are clearly separated.

### *G. Multi pass smearing*

For several images like the one shown in Figure 13, the separated image part may contain some text. This is possible mainly if some part of text is isolated and has a font much bigger than the rest of the text. For such cases, we repeat the run length smearing operation on the separated image. Figure 13 (a) is the original image. The text and image separated after application of our scheme are shown in Figure 13 (b) and (c) respectively. The image extracted in (c) shows some text content. When the smearing is applied on this image the text and image part is fully separated. This is shown in Figure 13 (d).

## Conclusions

Choice of the segmentation is very important step in the separation of newspaper and magazine document images. This choice is based on the quality of the input image, the required output quality and speed of processing. Run length smearing algorithm using boundary detection techniques are found to have good performance for the images having good separation between the text and the images pixels. Run length smearing algorithm and recursive segmentation technique are adopted in the segmentation stage to extract non-overlapping blocks embedded in the document.

The Enhanced CRLA has been chosen to build a number of document processing systems because of its advantages compared with other page segmentation techniques. For example, although the texture- analysis-based approaches are powerful to handle various page layouts, they are commonly time-consuming because of the pixel-level classification. The methods based on connected component analysis may have problems to extract large headline characters and thus they suit to deal with documents of certain character size. However, the original CRLA can only treat Manhattan page format; the present work extends its capability to cover both Manhattan and non-Manhattan layout and thus expands its application domain significantly. But the range of mean length of horizontal black runs (MBRL) and White-black transition count per unit width (MTC) in CRLA is not suitable for all type of document images.

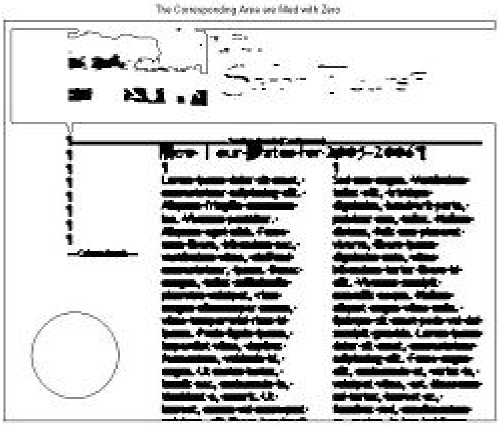

The Corresponding Area are filled with Zero
Safari Tours
New Tour Dates for 2005-2006¶

Figure 8. Smaller Area filled with Zero smearing

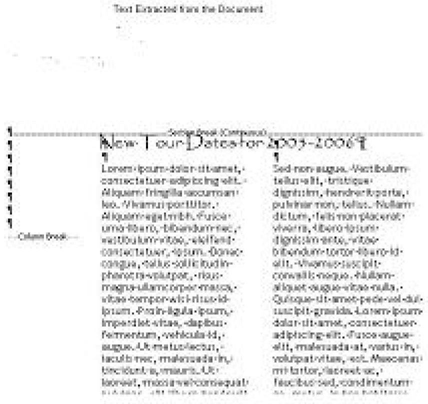

Text Extracted from the Document
Section Break (Continuous)
New Tour Dates for 2005-2006¶
Column Break

Figure 9. Text Extracted from given Document smearing

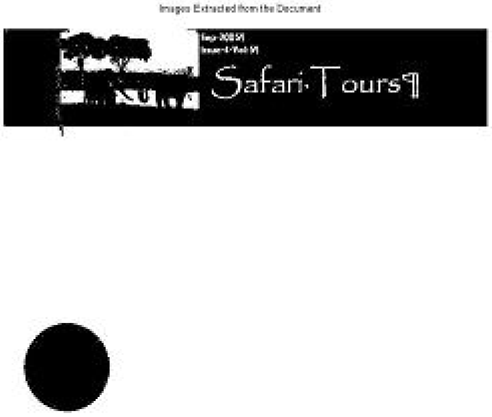


Figure 10. Image Extracted from given Document Image

To improve the performance of these techniques our method is implemented and it works well in case of all type of document images and works on text in several languages i.e.

© 2013 ACEEE
DOI: 01.IJSIP.4.1.1044

Hindi, English, Bangla, Telugu, Chinese. This method is language independent. If we see Projection profile method [1], [8] it is only applicable for Devanagari Document images. It needs that the layout of newspaper document image is very specific. Otherwise it will not separate the text-image region. Run length smearing using stripe merging method is only suitable when stripes formed during rum length smearing are perfect rectangle.

In our text-image separation system using boundary detection, there is no need of skew correction. Without skew correction it can work properly. Also, there is no requirement of labeling different sizes of text and images in the given documents. It can easily separate the text and image from document image. The method can be used recursively if the separated image contains some residue text. This system provides a good platform for implementing segmentation techniques suitable for developing real time processing techniques and processing large databases of newspaper and magazine document images.

## References

[1] Swapnil Khedekar Vemulapati Ramanaprasad Srirangaraj Setlur, “Text - Image Separation in Devanagari Documents”, Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR’03), 2003

[2] Q. Yuan, C.L. Tan, “Text Extraction from Gray Scale Document Images Using Edge Information”. Proceedings of the International Conference on Document Analysis and Recognition, ICDAR’01, Seattle, USA, pp. 302-306 September 10- 13, 2001

[3] A. Antonacopoulos and R T Ritchings “Segmentation and Classification of Document Images”, The Institution of Electrical Engineers, U.K. 1995

[4] [G. Nagy, S. Seth, “Hierarchical Representation of Optically Scanned Documents”. Proceedings of 7th Intl. Conf. on Pattern Recognition, Montreal, Canada, pp. 347-349, 1984.

[5] Kyong-Ho Lee, Student Yoon-Chul Choy, and Sung-Bae Cho, “ Geometric Structure Analysis of Document Images: A Knowledge-Based Approach”, IEEE transactions on Pattern Analysis and Machine Intelligence, Vol. 22, No. 11, November 2000.

[6] Jean Duong, Hubert Emptoz, “Features for Printed Document Image Analysis”, IEEE Transaction on Document Image Analysis Vol. 02, No. 17 November 2002.

[7] P. Mitchell, H. Yan, “Newspaper document analysis featuring connected line segmentation”, Proceedings of International Conference on on Document Analysis and Recognition, ICDAR’01, Seattle, USA, 2001.

[8] Abdel Wahab Zramdini and Rolf Ingold, “Optical font recognition from projection profiles”, Proceedings of Electronic Publishing. Vol. 6(3), pp.249–260, September, 1993.

[9] Kuo-Chin Fan, Chi-Hwa Liu, Yuan-Kai Wang, “Segmentation and classification of mixed text/graphics/image documents”, Pattern Recognition Letters, Elsevier, pp. 1201-1209, December 1994.

[10] Hung-Ming Sun, “Enhanced Constrained Run-Length Algorithm for Complex Layout Document Processing”, International Journal of Applied Science and Engineering pp. 297-309, 2006.

[11] Kuo-Chin Fan, Liang-Shen Wang, Yuan-Kai Wang , “Page segmentation and identification for intelligent signal processing”, Signal Processing, Elsevier, pp. 329-346, 1995.

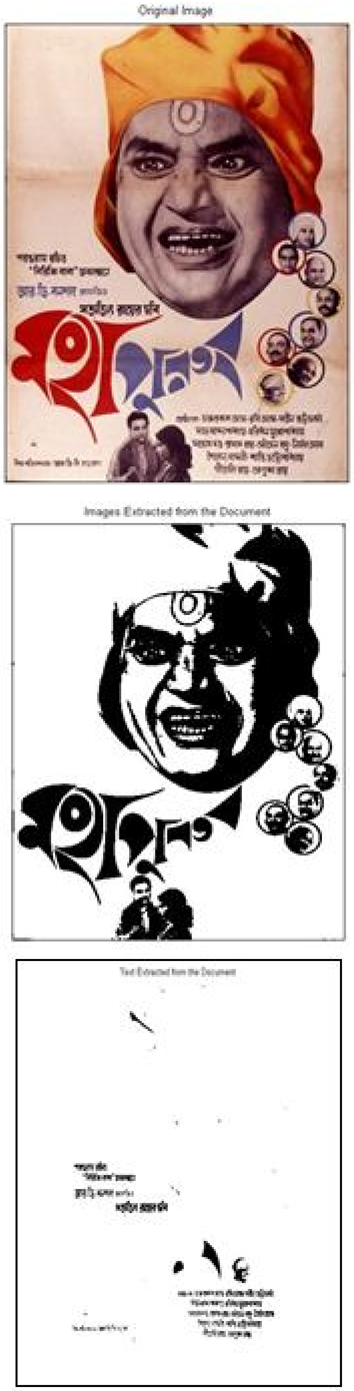

Original Image
Images Extracted from the Document
Text Extracted from the Document

Figure 11. Text image separated for sample 2

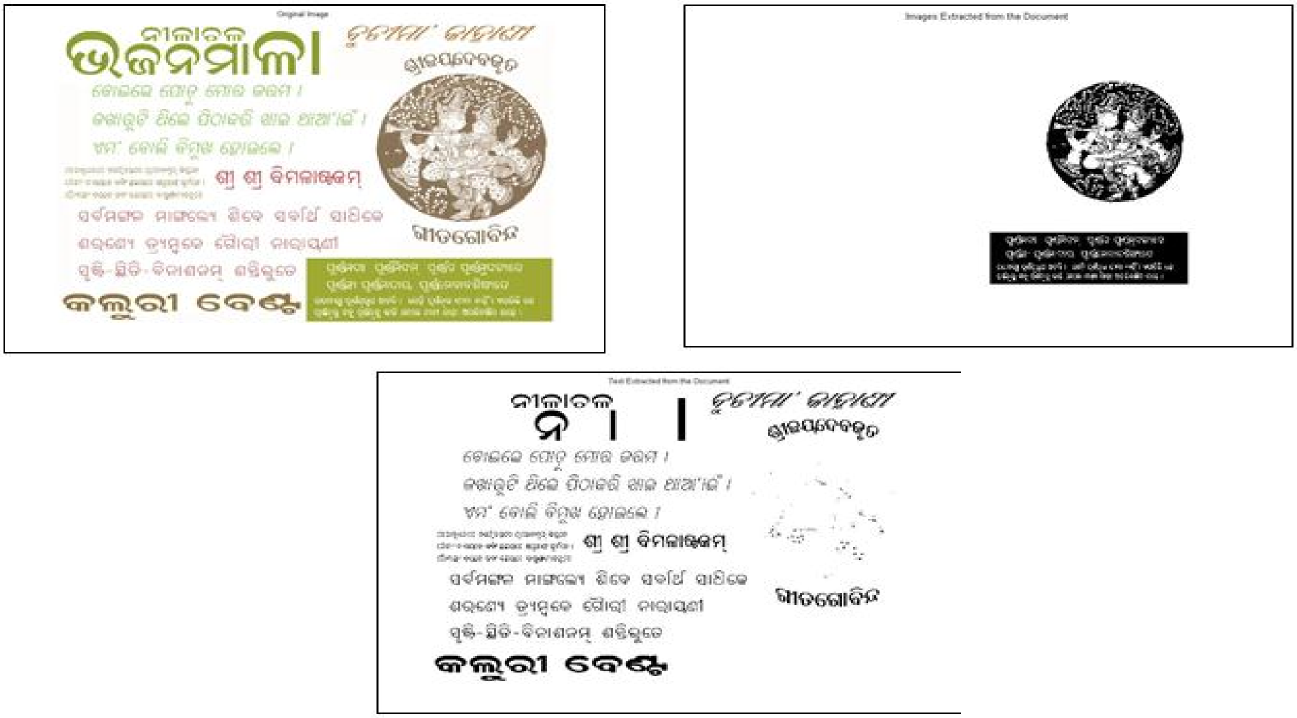
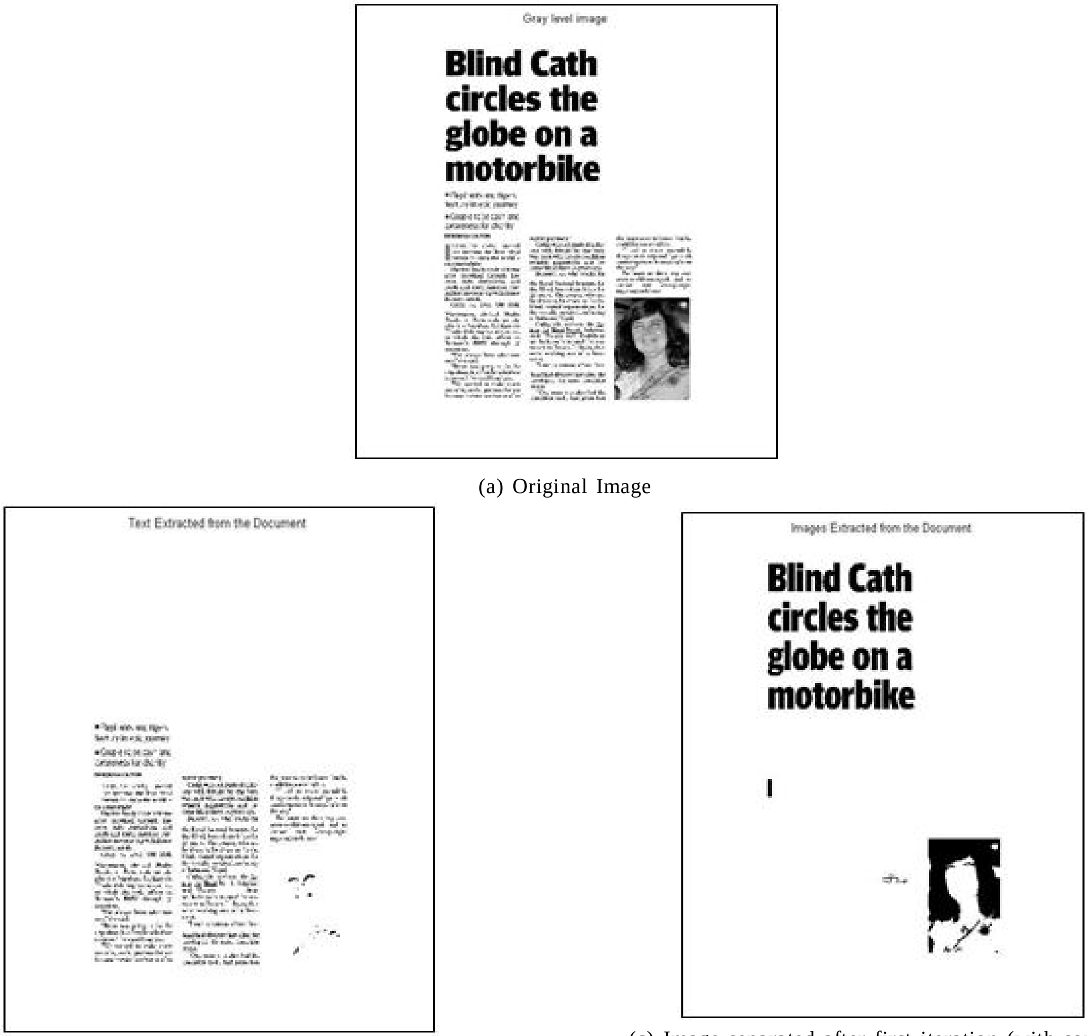

Original Image
Images Extracted from the Document
Text Extracted from the Document

Figure 12. Text image separated for sample 3

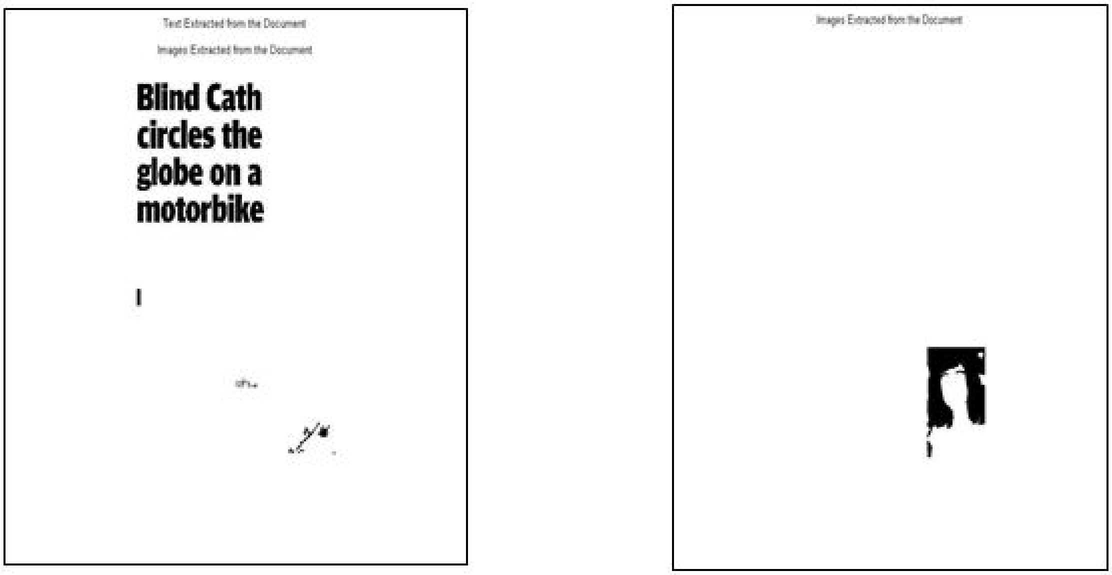

Gray level image
Blind Cath circles the globe on a motorbike
Text Extracted from the Document
Images Extracted from the Document
Blind Cath circles the globe on a motorbike

(a) Original Image

(b) Text separated after first iteration

(c) Image separated after first iteration (with some text)

© 2013 ACEEE
DOI: 01.IJSIP.4.1.1044


Text Extracted from the Document
Images Extracted from the Document
Blind Cath circles the globe on a motorbike
Images Extracted from the Document

(d) Text and image separation after second iteration of smearing on image in (c)

Figure 13. Text separation requiring two passes

DOI: 01.IJSIP.4.1.1044